

A review of undirected and acyclic directed Gaussian Markov model selection and estimation

Irene Córdoba-Sánchez
Concha Bielza
Pedro Larrañaga

Abstract

Markov models lie at the interface between statistical independence in a probability distribution and graph separation properties. We review model selection and estimation in directed and undirected Markov models with Gaussian parametrization, emphasizing the main similarities and differences. These two model types are foundationally similar but not equivalent, as we highlight. We report existing results with a unified notation and terminology, taking into account literature from both the artificial intelligence and statistics research communities, which first developed these models. Finally, we point out the main active research areas and open problems now existing with regard to these traditional, albeit rich, Markov models.

Keywords. Conditional independence, Gaussian Bayesian network, Gaussian graphical model, Gaussian Markov model, model selection, parameter estimation

1 Introduction

Markov models, or probabilistic graphical models, explicitly establish a correspondence between statistical independence in a probability distribution and certain separation criteria holding in a graph. They originated at the interface between statistics, dominated by Markov random fields (Darroch et al., 1980), and artificial intelligence, with a focus on Bayesian networks (Pearl, 1985, 1986). Markov random fields and Bayesian networks are now both considered traditional models. Nevertheless, they are still widely applied and attract a significant amount of research nowadays (Daly et al., 2011; Wermuth, 2015). A feature that they have in common is that they both model conditional independence: Bayesian networks relate conditional independence to acyclic directed graphs, whereas it is associated with undirected graphs in Markov fields.

Markov random fields (Grimmett, 1973) are the oldest type of Markov models. They were preceded only by special cases such as the Ising model (Isham, 1981) or Markov chains. Markov random fields generalized the correspondence between Gibbs measures (Besag, 1974) and Markov properties. The term *graphical model* was not introduced for them until Darroch et al. (1980) linked the graphical ideas for contingency tables with Markov properties of discrete Markov fields. They were also termed *Markov networks* (Pearl, 1988) by artificial intelligence researchers by analogy with Bayesian networks. In contrast, acyclic digraph models became popular after the introduction of *influence diagrams* for decision-making processes by Howard and Matheson in 1981 (article reprinted in Howard and Matheson (2005)). The probabilistic reduction of influence diagrams was studied at length by Pearl (Pearl, 1988), who renamed probabilistic influence diagrams as *Bayesian networks* or *influence networks* (Pearl, 1985). Furthermore, acyclic directed

Markov models were also called *directed Markov fields* (Lauritzen et al., 1990), by analogy with Markov random fields.

In this paper, we review the existing methods for model selection and estimation in undirected and acyclic directed Markov models with Gaussian parametrization. The multivariate Gaussian distribution is one of the most widely developed and applied statistical families in this context (Werhli et al., 2006; Ibáñez et al., 2016). It provides for an explicit parametric comparison of their similarities and differences. As we have seen, their terminology varies considerably due to the highly interdisciplinary nature of these Markov models. This issue has further ramifications in the Gaussian case. Gaussian acyclic directed models were implicitly studied by Wermuth (1980) as *linear recursive regression systems*. Furthermore, geneticist Sewall Wright in 1918 developed the method of *path coefficients* (Wright, 1934), nowadays known as *path analysis*, where linearly related variables were represented using a directed acyclic graph. In the undirected case, we find a similar situation. Dempster (1972) suggested estimating the inverse of the covariance matrix (concentration matrix) by assuming certain entries to be equal to zero. Interestingly, although Dempster did not have any graphical interpretation in mind, such zero entries are directly associated with missing edges in undirected Gaussian Markov models, which is why even nowadays they are sometimes still called *covariance selection* models. Wermuth (1976) was the first to note such correspondence, by analogies with contingency tables in the discrete case. Furthermore, in further developments of Markov models, Cox and Wermuth (1993) referred to these models as *concentration graphs*, resembling the zero entries in the concentration matrix. In this paper, we use a unified notation and keep it throughout all the sections, allowing for a direct comparison within and between the different methods and foundational aspects reviewed. Figure 1 is a timeline illustrating the origins of Markov models above presented.

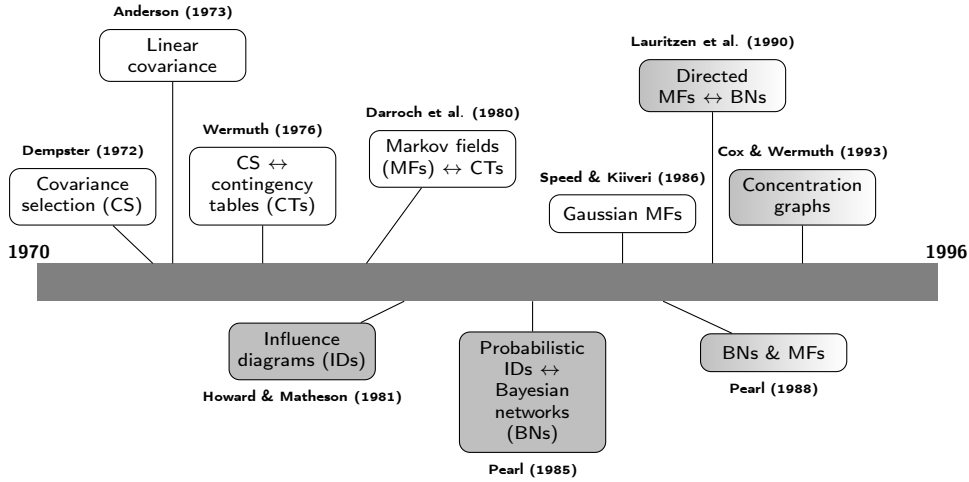


Figure 1: Timeline of the origins of Gaussian Markov models. Papers from the statistical community appear above the line and papers from other research areas are shown below the line. Thematically, grey filled boxes are papers about acyclic directed Markov models, white boxes refer to undirected models, and gradient filled boxes deal with both types of models.

We begin the paper by introducing undirected and acyclic directed Markov models distribution-free assumptions in Section 2. Many foundational relationships can already be established from this general perspective, as we show. Next, in Section 3, we explore the consequences of restricting the models to multivariate Gaussian distributions, exploring how this helps to simplify the

model description in both cases. Maximum likelihood estimation under a previously selected model, that is, under a fixed graph, is then reviewed in Section 4. Procedures for model selection via hypothesis testing are reported in Section 5. When maximum likelihood estimators are not guaranteed to exist, a popular technique is to employ regularisation, which we outline in Section 6. Finally, we discuss the alternative Bayesian approach for Markov models in Section 7. We conclude the paper with a discussion of the current main active lines of research and open problems with respect to each of the reviewed areas in Section 8. An explanation of basic concepts from graph theory used throughout the paper is provided in Appendix A.

2 Undirected and acyclic directed Markov models

The Markov models that we review associate conditional independence in random vectors $\mathbf{X} = (X_1, \dots, X_p)^t$ with undirected graph and acyclic digraph separation properties. This is explicitly specified by the *Markov properties* of the distribution of \mathbf{X} , which are in turn based on what are known as *independence relations*.

In the following, for arbitrary $I \subseteq \{1, \dots, p\}$, we will denote the $|I|$ -dimensional subvector of \mathbf{X} by $\mathbf{X}_I := (X_i)_{i \in I}$. Conditional independence will be expressed as $\mathbf{X}_I \perp\!\!\!\perp \mathbf{X}_J \mid \mathbf{X}_K$, which represents the statement ‘ \mathbf{X}_I is conditionally independent from \mathbf{X}_J given \mathbf{X}_K ’ (see Dawid (1979)). $\mathcal{G} = (V, E)$ will denote a graph with vertex set V and edge set E , and for $U \subseteq V$, \mathcal{G}_U will be the subgraph induced by U . When the graph is acyclic directed, it will usually (although not always) be denoted as \mathcal{D} . Its undirected version or skeleton is \mathcal{D}^U , and its moral graph \mathcal{D}^m . For $v \in V$ in an acyclic directed graph, its parent, non-descendant and predecessor sets will be denoted as $\text{pa}(v)$, $\text{nd}(v)$ and $\text{pr}(v)$, respectively; for $U \subseteq V$, its ancestral set will be $\text{An}(U)$. Analogously, if $v \in V$ in an undirected graph, the neighbour and closure sets of v will be $\text{ne}(v)$ and $\text{cl}(v)$, respectively. See Appendix A for a brief introduction to graph theory where the above concepts are further explained.

2.1 Independence relations

An *independence relation* over a set $V = \{1, \dots, p\}$ is a collection \mathcal{I} of triples (A, B, C) , where A , B and C are pairwise disjoint subsets of V . It is called a *semi-graphoid* when the following conditions are met:

- if $(A, B, C) \in \mathcal{I}$ then $(B, A, C) \in \mathcal{I}$,
- if $(A, B \cup C, D) \in \mathcal{I}$ then $(A, C, D) \in \mathcal{I}$ and $(A, B, C \cup D) \in \mathcal{I}$,
- if $(A, B, C \cup D) \in \mathcal{I}$ and $(A, C, D) \in \mathcal{I}$ then $(A, B \cup C, D) \in \mathcal{I}$,

and a *graphoid* when, additionally, if $(A, B, C \cup D) \in \mathcal{I}$ and $(A, C, B \cup D) \in \mathcal{I}$ then $(A, B \cup C, D) \in \mathcal{I}$ (Pearl and Paz, 1987).

Independence relations occur in different contexts that are relevant for Markov models. Specifically, an independence relation \mathcal{I} over $V = \{1, \dots, p\}$ is said to be *induced* by

- an undirected graph $\mathcal{G} = (V, E)$ if $(A, B, S) \in \mathcal{I} \iff A$ and B are separated by S in \mathcal{G} ,
- an acyclic digraph $\mathcal{D} = (V, A)$ if $(A, B, S) \in \mathcal{I} \iff A$ and B are separated by S in $(\mathcal{D}_{\text{An}(A \cup B \cup S)})^m$,
- a p -dimensional random vector \mathbf{X} if $(A, B, S) \in \mathcal{I} \iff \mathbf{X}_A \perp\!\!\!\perp \mathbf{X}_B \mid \mathbf{X}_S$.

Graph-induced independence relations are always graphoids, while probabilistic independence relations are always semi-graphoids. Additional assumptions on the probability spaces involved are required for probabilistic independence relations to be graphoids (Dawid, 1980).

The core of Markov models is the relationship between induced independence relations, which we will denote by $\mathcal{I}(\cdot)$ with the argument being the inducing element. Specifically, if \mathcal{G} is an undirected (acyclic directed) graph, an undirected (directed) *Markov model* is defined as

$$\mathcal{M}(\mathcal{G}) := \{P_{\mathbf{X}} : \mathcal{I}(\mathcal{G}) \subseteq \mathcal{I}(\mathbf{X})\},$$

where the random vectors \mathbf{X} are defined over the same probability space and $P_{\mathbf{X}}$ denotes their distribution. These models are non-empty (Geiger and Pearl, 1990, 1993), that is, for any undirected or acyclic directed graph, there is always a probability distribution whose independence model contains the one induced by the graph.

2.2 Markov properties

When a distribution $P_{\mathbf{X}}$ belongs to $\mathcal{M}(\mathcal{G})$ for an undirected or acyclic directed graph \mathcal{G} , it is said that $P_{\mathbf{X}}$ is *globally \mathcal{G} -Markov* or satisfies the *global Markov property* with respect to \mathcal{G} . Other weaker Markov properties can be defined that are usually capable of simplifying the model. Specifically, if $\mathcal{G} = (V, E)$ is an undirected graph, then the probability distribution $P_{\mathbf{X}}$ of \mathbf{X} is said to be

- *pairwise \mathcal{G} -Markov* if $X_u \perp\!\!\!\perp X_v \mid \mathbf{X}_{V \setminus \{u, v\}}$ for all $uv \notin E$,
- *locally \mathcal{G} -Markov* if $X_v \perp\!\!\!\perp \mathbf{X}_{V \setminus \text{cl}(v)} \mid \mathbf{X}_{\text{ne}(v)}$ for all $v \in V$,

whereas if \mathcal{G} is an acyclic digraph, then $P_{\mathbf{X}}$ is called

- *pairwise \mathcal{G} -Markov* if $X_u \perp\!\!\!\perp X_v \mid \mathbf{X}_{\text{nd}(u) \setminus \{v\}}$ for all $u \in V, v \in \text{nd}(u) \setminus \text{pa}(u)$,
- *locally \mathcal{G} -Markov* if $X_v \perp\!\!\!\perp \mathbf{X}_{\text{nd}(v) \setminus \text{pa}(v)} \mid \mathbf{X}_{\text{pa}(v)}$ for all $v \in V$.

The three Markov properties are equivalent when \mathcal{G} is acyclic directed (Lauritzen et al., 1990), while, if \mathcal{G} is undirected, this equivalence is only guaranteed when $\mathcal{I}(\mathbf{X})$ is a graphoid (Pearl, 1988). Suffice it for $P_{\mathbf{X}}$ to admit a continuous and strictly positive density for this to occur. This sufficient condition was obtained in different forms by several authors, but it is usually attributed to Hammersley and Clifford (1971), who were the first to outline the proof for the discrete case (Speed, 1979). It relies on an additional characterization of a probability distribution with respect to \mathcal{G} : if $\mathfrak{C}(\mathcal{G})$ denotes the class of cliques of \mathcal{G} , the density function f of $P_{\mathbf{X}}$ is said to *factorize according to \mathcal{G}* when there exists a set $\{\psi_C(\mathbf{x}_C) : C \in \mathfrak{C}(\mathcal{G}), \psi_C \geq 0\}$ such that

$$f(\mathbf{x}) = \prod_{C \in \mathfrak{C}(\mathcal{G})} \psi_C(\mathbf{x}_C). \quad (1)$$

When Equation (1) holds, then $P_{\mathbf{X}}$ is globally \mathcal{G} -Markov, while if f is continuous and strictly positive, the pairwise Markov property implies Equation (1). This yields the equivalence of Markov properties.

Finally, recall that the nodes of an acyclic digraph $\mathcal{D} = (V, A)$ can be totally ordered such that if $(u, v) \in A$, then $u \in \text{pr}(v)$. This gives rise to another Markov property, exclusive to acyclic digraphs: $P_{\mathbf{X}}$ is said to be *ordered \mathcal{D} -Markov* if $X_v \perp\!\!\!\perp \mathbf{X}_{\text{pr}(v) \setminus \text{pa}(v)} \mid \mathbf{X}_{\text{pa}(v)}$ for all $v \in V$. This property is also equivalent to the global, local and pairwise Markov properties (Lauritzen et al., 1990).

2.3 Independence and Markov equivalence

When the Markov models defined by two graphs \mathcal{G} and $\tilde{\mathcal{G}}$, with the same vertex set V , coincide, such graphs are said to be *Markov equivalent*. A simpler notion, also implying Markov equivalence, is *independence equivalence*, holding when $\mathcal{I}(\mathcal{G}) = \mathcal{I}(\tilde{\mathcal{G}})$. Independence equivalence is implied by Markov equivalence under fairly general circumstances (Studený, 2005, §6.1), which is why most authors treat them as the same notion. The best suited graph for the Markov model can be chosen based on these equivalences.

We will first characterize equivalence within undirected graphs. Since there exists a unique edge-minimal undirected graph \mathcal{G} such that $\mathcal{I}(\mathcal{G}) \subseteq \mathcal{I}$ for each graphoid \mathcal{I} over V (Pearl and Paz, 1987), it follows that $\mathcal{I}(\mathcal{G}) = \mathcal{I}(\tilde{\mathcal{G}})$ (independence equivalence) if and only if \mathcal{G} and $\tilde{\mathcal{G}}$ are identical. Furthermore, if we assume that $\mathcal{I}(\mathbf{X})$ is a graphoid for all $P_{\mathbf{X}} \in \mathcal{M}(\mathcal{G})$, then a unique edge-minimal $\tilde{\mathcal{G}}$ exists, with $\tilde{\mathcal{G}} \subseteq \mathcal{G}$, such that $\mathcal{M}(\mathcal{G}) = \mathcal{M}(\tilde{\mathcal{G}})$ (Markov equivalence), that is, a unique undirected graph can be chosen as representative of each undirected Markov model.

In contrast, acyclic digraphs are not, generally, unique representations of a Markov model, since $\mathcal{I}(\mathcal{D}) = \mathcal{I}(\tilde{\mathcal{D}})$ if and only if \mathcal{D} and $\tilde{\mathcal{D}}$ have the same skeleton and the same v-structures (Verma and Pearl, 1991). However, unique representatives can be constructed: Let \mathfrak{D}_p be the set of acyclic digraphs over $V = \{1, \dots, p\}$ and define an equivalence relation \sim in \mathfrak{D}_p as $\mathcal{D} \sim \tilde{\mathcal{D}} \iff \mathcal{I}(\mathcal{D}) = \mathcal{I}(\tilde{\mathcal{D}})$. The quotient space of \sim is $\mathfrak{D}_p / \sim = \{[\mathcal{D}] : \mathcal{D} \in \mathfrak{D}_p\}$, where $[\mathcal{D}] := \{\tilde{\mathcal{D}} \in \mathfrak{D}_p : \tilde{\mathcal{D}} \sim \mathcal{D}\}$ is the *Markov equivalence class*; thus, $\mathcal{M}(\tilde{\mathcal{D}}) = \mathcal{M}(\mathcal{D})$ for all $\tilde{\mathcal{D}} \in [\mathcal{D}]$, that is, $[\mathcal{D}]$ is the unique representative of the directed Markov model.

Equivalence between acyclic directed and undirected graphs was first obtained by Wermuth (1980) for multivariate Gaussian distributions and by Wermuth and Lauritzen (1983) for contingency tables, and then generalized by Frydenberg (1990) for graphoid-inducing distributions. When \mathcal{G} is an undirected graph, $\mathcal{M}(\mathcal{G}) = \mathcal{M}(\mathcal{D})$ for some acyclic digraph \mathcal{D} if and only if \mathcal{G} is chordal. Conversely, an acyclic digraph \mathcal{D} is Markov equivalent to its skeleton \mathcal{D}^U if and only if \mathcal{D} contains no v-structures. Furthermore, \mathcal{D} can be related to its moral graph. This requires an analogous formula to Equation (1): a density function f is said to *recursively factorize* according to \mathcal{D} when

$$f(\mathbf{x}) = \prod_{v \in V} f(x_v \mid \mathbf{x}_{\text{pa}(v)}).$$

This characterization is equivalent to the Markov properties, and also implies that f factorizes as in Equation (1) with respect to the moral graph \mathcal{D}^m (Lauritzen et al., 1990). This means that $P_{\mathbf{X}}$ is always globally \mathcal{D}^m -Markov for continuous \mathbf{X} , and thus $\mathcal{M}(\mathcal{D}) \subseteq \mathcal{M}(\mathcal{D}^m)$, with the equality only holding when $\mathcal{D}^m = \mathcal{D}^U$.

Example 1. Figure 2 illustrates the above concepts. The graph in Figure 2a is not chordal, and thus there is no Markov equivalent acyclic digraph. Figure 2b is a chordal cover of Figure 2a, and a Markov equivalent orientation is depicted in Figure 2c. The acyclic digraph in Figure 2d has v-structures, emphasized in dark grey, and thus cannot be Markov equivalent to its skeleton (Figure 2a). The moral graph in Figure 2d is Figure 2e, which in fact is another chordal cover of Figure 2a, and thus none of its orientations will be Markov equivalent to Figure 2c.

There is certain symmetry in the foundational properties of both undirected and acyclic directed Markov models, that we have presented throughout this section. To make their relationship more explicit, we have summarized such results in Table 1.

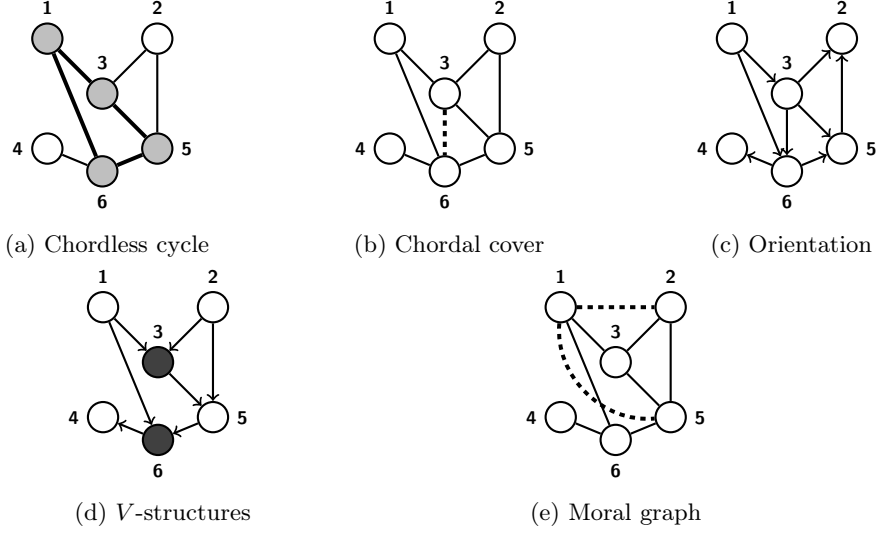


Figure 2: Markov equivalence.

	Undirected (\mathcal{G})	Directed (\mathcal{D})
$\mathcal{I}(\cdot)$ properties	Graphoid	Graphoid
Defining MP	$\mathbf{X}_A \perp\!\!\!\perp \mathbf{X}_B \mid \mathbf{X}_S$ for A and B separated by S in \mathcal{G}	$\mathbf{X}_A \perp\!\!\!\perp \mathbf{X}_B \mid \mathbf{X}_S$ for A and B separated by S in $(\mathcal{D}_{\text{An}(A \cup B \cup S)})^m$
Factorization	$\prod_{C \in \mathcal{C}(\mathcal{G})} \psi_C(\mathbf{x}_C)$	$\prod_{v \in V} f(x_v \mid \mathbf{x}_{\text{pa}(v)})$
MPs equivalent	If $\mathcal{I}(\mathbf{X})$ is a graphoid for all $P_{\mathbf{X}}$ in $\mathcal{M}(\mathcal{G})$	Always
Uniqueness	Yes	No (Markov equivalence classes)
ME graph	Iff \mathcal{G} is chordal: an orientation of \mathcal{G}	Iff \mathcal{D} has no v-structures: \mathcal{D}^U

Table 1: Summary of properties for Markov models, with the following abbreviations: MP for Markov property, ME for Markov equivalent.

3 Gaussian parametrization

When restricting to multivariate Gaussian distributions, further symmetry can be found between undirected and acyclic directed Markov models: in both cases, we find connections between conditional independence and vanishing numerical parameters in the underlying Gaussian distribution.

In the following, the elements of a real $q \times r$ matrix $\mathbf{M} \in \mathbb{M}_{q \times r}(\mathbb{R})$ will be denoted by m_{ij} , where $i \in \{1, \dots, q\}$ and $j \in \{1, \dots, r\}$. \mathbf{M}_{IJ} will be the $|I| \times |J|$ submatrix of \mathbf{M} , where $I \subseteq \{1, \dots, q\}$ and $J \subseteq \{1, \dots, r\}$, and we will use \mathbf{M}_{IJ}^{-1} as $(\mathbf{M}_{IJ})^{-1}$. $\mathbb{S}^{>0}$ and $\mathbb{S}^{\geq 0}$ will represent the sets of positive and semi-positive definite symmetric matrices, respectively. The p -variate Gaussian distribution is denoted by $\mathcal{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, where $\boldsymbol{\mu} \in \mathbb{R}^p$ is the mean vector and $\boldsymbol{\Sigma} \in \mathbb{S}^{>0}$ is the covariance matrix. \mathbf{I}_p will denote the $p \times p$ identity matrix. Dimensional subscripts will often be dropped if the dimension of the respective object is clear from the context.

3.1 Conditional independence and the multivariate Gaussian distribution

Let $V = \{1, \dots, p\}$, and consider random vectors \mathbf{X} distributed as $\mathcal{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. Let the concentration matrix of \mathbf{X} be $\boldsymbol{\Omega} = \boldsymbol{\Sigma}^{-1}$, with elements ω_{uv} for $u, v \in V$. Basic facts from multivariate Gaussian theory are that for $i, j \in V$, $X_i \perp\!\!\!\perp X_j$ is equivalent to $\sigma_{ij} = 0$, and if (I, J) is a partition of V , then $\mathbf{X}_I | \mathbf{x}_J$ is distributed as $\mathcal{N}_{|I|}(\boldsymbol{\mu}_I, \boldsymbol{\Sigma}_{I|J})$, where $\boldsymbol{\Sigma}_{I|J} = \boldsymbol{\Sigma}_{II} - \boldsymbol{\Sigma}_{IJ} \boldsymbol{\Sigma}_{JJ}^{-1} \boldsymbol{\Sigma}_{JI}$ (Anderson, 2003). Thus, for $i, k \in I$, we have that $X_i \perp\!\!\!\perp X_k | \mathbf{x}_J$ is equivalent to $\sigma_{ik|J} = 0$, the (i, k) element in the conditional covariance matrix $\boldsymbol{\Sigma}_{I|J}$.

A correspondence can be established between those zeros in $\boldsymbol{\Sigma}_{I|J}$ and zero patterns in other, more representative matrices (Wermuth, 1976, 1980) as follows. The matrix $\boldsymbol{\Sigma}_{IJ} \boldsymbol{\Sigma}_{JJ}^{-1}$ is usually denoted by $\mathbf{B}_{I|J}$ and called the matrix of *regression coefficients* of \mathbf{X}_I on \mathbf{X}_J . If we further let $\boldsymbol{\Omega}_{I|J} := \boldsymbol{\Sigma}_{I|J}^{-1}$, then standard theory for partitioned matrices (Horn and Johnson, 2012) gives

$$\boldsymbol{\Sigma}_{I|J} = \boldsymbol{\Omega}_{II}^{-1}, \quad (2)$$

$$\mathbf{B}_{I|J} = -\boldsymbol{\Omega}_{II}^{-1} \boldsymbol{\Omega}_{IJ}. \quad (3)$$

Using now the fact that $X_i \perp\!\!\!\perp X_k | \mathbf{x}_J$ is equivalent to $\sigma_{ik|J} = 0$ in a partition (I, J) of V , conditional independence can be related with $\boldsymbol{\Omega}$ and $\mathbf{B}_{I|J}$ via $\boldsymbol{\Sigma}_{I|J}$: from Equation (2) we get, for $i, k \in V$,

$$X_i \perp\!\!\!\perp X_k | \mathbf{X}_{V \setminus \{i, k\}} \iff \omega_{ik} = 0, \quad (4)$$

whereas from Equation (3) it follows that, for $J \subseteq V$, $i, k \in V \setminus J$,

$$X_i \perp\!\!\!\perp X_k | \mathbf{X}_J \iff \beta_{ik|J \cup \{k\}} = 0, \quad (5)$$

where $\beta_{ik|J \cup \{k\}}$ is the v entry in the vector $\boldsymbol{\beta}_{i|J \cup \{k\}}^t$, that is, the coefficient of X_k on the regression of X_i on $\mathbf{x}_{J \cup \{k\}}$. The original notation for this, introduced by Yule (1907), was $\beta_{ik \cdot J}$, that is, k is implicitly considered as included in the conditioning indexes. However, we have opted for the alternative explicit notation $\beta_{ik|J \cup \{k\}}$, since it simplifies the notation in later sections.

3.2 Gaussian Markov models

In the Gaussian case, undirected Markov models are in correspondence with the concentration matrix (Equation (4)), while for acyclic digraphs this correspondence is with the regression coefficients (Equation (5)), as we will now show.

Consider a Gaussian distribution $P_{\mathbf{X}}$ pairwise \mathcal{G} -Markov with respect to some undirected graph \mathcal{G} . Letting $\mathbb{S}^{\succ 0}(\mathcal{G}) := \{\mathbf{M} \in \mathbb{S}^{\succ 0} : m_{uv} = 0 \text{ for all } uv \notin E\}$ and noting that in a multivariate Gaussian distribution the three Markov properties are equivalent, we obtain from Equation (4) that $\boldsymbol{\Omega} \in \mathbb{S}^{\succ 0}(\mathcal{G})$ if and only if $P_{\mathbf{X}} \in \mathcal{M}(\mathcal{G})$. Hence, we can characterize the *Gaussian undirected Markov model* as

$$\mathcal{N}(\mathcal{G}) = \{\mathcal{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}) : \boldsymbol{\Sigma}^{-1} \in \mathbb{S}^{\succ 0}(\mathcal{G}), \boldsymbol{\mu} \in \mathbb{R}^p\}. \quad (6)$$

In the directed case, the redefinition is not so direct. Let $\mathcal{D} = (V, A)$ be an acyclic digraph, and assume, for notational simplicity, that its node set is already ancestrally ordered as $1 \preceq \dots \preceq p$. If $P_{\mathbf{X}}$ is ordered \mathcal{D} -Markov, whenever $v \in \text{pr}(u) \setminus \text{pa}(u)$, we have $X_u \perp\!\!\!\perp X_v | \mathbf{X}_{\text{pa}(u)}$, which in the Gaussian case is equivalent to $\beta_{uv|\text{pa}(u) \cup \{v\}} = 0$, as in Equation (5). Furthermore, the ancestral order gives $\beta_{uv|\text{pa}(u) \cup \{v\}} = \beta_{uv|\text{pr}(u)}$, hence $P_{\mathbf{X}} \in \mathcal{M}(\mathcal{D})$ if and only if $\beta_{uv|\text{pr}(u)} = 0$ for all $u \in V$, $v \in \text{pr}(u) \setminus \text{pa}(u)$.

The above requirement on the regression coefficients can be expressed in matrix notation as follows. Let \mathbf{B} be the triangular matrix defined as $b_{uv} = 0$ for $v < u$, $v \notin \text{pa}(u)$, and $b_{uv} = \beta_{uv|\text{pr}(u)}$ otherwise. Then we have the linear regression equation

$$\mathbf{X} = \boldsymbol{\mu} + \mathbf{B}(\mathbf{X} - \boldsymbol{\mu}) + \mathbf{E}, \quad (7)$$

where $\mathbf{E} \sim \mathcal{N}_p(0, \mathbf{V})$ with \mathbf{V} the diagonal matrix of conditional variances $\sigma_{uu|\text{pr}(u)}$. We can rearrange Equation (7) as $\mathbf{X} = \mathbf{U}^{-1}\boldsymbol{\xi} + \mathbf{U}^{-1}\mathbf{E}$, with $\boldsymbol{\xi} := \mathbf{U}\boldsymbol{\mu}$ and $\mathbf{U} := \mathbf{I}_p - \mathbf{B}$. $(\boldsymbol{\xi}, \mathbf{B}, \mathbf{V})$ are sometimes called the \mathcal{D} -parameters of $(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ (Andersson and Perlman, 1998), since they allow to recover $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$. In fact, $\boldsymbol{\Sigma}$ can be decomposed as $\boldsymbol{\Sigma} = \mathbf{U}^{-1}\mathbf{V}\mathbf{U}^{-t}$, and this decomposition uniquely determines $\boldsymbol{\Sigma}$ from \mathbf{U} and \mathbf{V} (Horn and Johnson, 2012). Hence, if we define $\mathbb{M}(\mathcal{D}) := \{\mathbf{M} \in \mathbb{M}_{p \times p}(\mathbb{R}) : m_{uv} = 0 \text{ for all } (u, v) \notin A\}$ and the set $\boldsymbol{\Delta}_p$ of $p \times p$ diagonal matrices, we can characterize the *Gaussian directed Markov model* as

$$\mathcal{N}(\mathcal{D}) = \{\mathcal{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}) : \boldsymbol{\Sigma}^{-1} \in \mathbb{S}^{>0}(\mathcal{D}), \boldsymbol{\mu} \in \mathbb{R}^p\}, \quad (8)$$

where $\mathbb{S}^{>0}(\mathcal{D}) := \{\mathbf{M} \in \mathbb{S}^{>0} : \mathbf{M} = \mathbf{U}^t \mathbf{V}^{-1} \mathbf{U}, \mathbf{I}_p - \mathbf{U} \in \mathbb{M}(\mathcal{D}), \mathbf{V} \in \boldsymbol{\Delta}_p\}$.

4 Maximum likelihood estimation

Exponential family theory simplifies maximum likelihood estimation. In a regular exponential family $\mathcal{F}_{\mathcal{H}}$ with canonical parameter $\boldsymbol{\eta}$ over \mathcal{H} and sufficient statistic \mathbf{t} , the maximum of the likelihood function given a random sample $\mathbf{X} = \mathbf{x}$ is reached in \mathcal{H} if and only if $\mathbf{t}(\mathbf{x})$ belongs to the interior of $\mathcal{C}(\mathbf{t})$, the closed convex hull of the support of the distribution of \mathbf{t} , denoted by $\text{int}(\mathcal{C}(\mathbf{t}))$. When this occurs, the maximum is the only $\boldsymbol{\eta} \in \mathcal{H}$ satisfying $\mathbb{E}[\mathbf{t}(\mathbf{X})] = \mathbf{t}(\mathbf{x})$.

A random sample $\mathbf{X} = \mathbf{x}$ from a multivariate Gaussian $\mathcal{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, where \mathbf{x} has dimension $p \times N$, is also a regular exponential family, with canonical parameter $\boldsymbol{\eta} = (\boldsymbol{\Omega}\boldsymbol{\mu}, -\boldsymbol{\Omega}/2)$. The sufficient statistic for the random sample is $\mathbf{t}(\mathbf{X}) = (N\bar{\mathbf{X}}, \mathbf{X}\mathbf{X}^t)$ with $N\bar{\mathbf{X}} = \sum_{n=1}^N \mathbf{X}^{(n)}$, and its convex support is $\mathcal{C}(\mathbf{t}) = \{(\mathbf{v}, \mathbf{M}) \in \mathbb{R}^p \times \mathbb{S}_p : \mathbf{M} - \mathbf{v}\mathbf{v}^t/N \in \mathbb{S}^{\geq 0}\}$. Hence, the maximum likelihood estimator for $(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ in the unrestricted case (that is, when not assuming a Markov model) exists if and only if $\mathbf{Q} := \mathbf{x}\mathbf{x}^t - N\bar{\mathbf{x}}\bar{\mathbf{x}}^t \in \mathbb{S}^{>0}$. When it exists, the solution is $(\bar{\mathbf{x}}, \mathbf{Q}/N)$. When the model is Markov for some graph \mathcal{G} , then we are in presence of a subfamily of this regular exponential family, with different properties depending on whether \mathcal{G} is undirected or acyclic directed, as we will show.

If \mathcal{G} is an undirected graph, $\mathbb{R}^p \times \mathbb{S}^{>0}(\mathcal{G})$ is an affine subspace of $\mathbb{R}^p \times \mathbb{S}^{>0}$, and thus $\mathcal{N}(\mathcal{G})$ is also a regular exponential family (Barndorff-Nielsen, 1978). Maximum likelihood estimators are thus obtained as happened in the unrestricted case. In particular, it is usually assumed that all distributions in $\mathcal{N}(\mathcal{G})$ have zero mean, since whether $\mathcal{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ belongs to $\mathcal{N}(\mathcal{G})$ depends only on $\boldsymbol{\Sigma}$ (Equation 8). Let $\mathbf{Q} := \mathbf{x}\mathbf{x}^t$ and $\mathbf{Q}^{\mathcal{G}}$ be the projection of \mathbf{Q} on $E \cup \{uu : u \in V\}$, that is, such that $q_{uv}^{\mathcal{G}} = 0$ for all $uv \notin E$ with $u \neq v$. Then, the sufficient statistic for $\mathcal{N}(\mathcal{G})$ is $\mathbf{t}(\mathbf{x}) = \mathbf{Q}^{\mathcal{G}}$ (Lauritzen, 1996). Its convex support is $\mathcal{C}(\mathbf{t}) = \{\mathbf{P}^{\mathcal{G}} : \mathbf{P} \in \mathbb{S}^{\geq 0}\}$, alternatively referred to as the set of projections extensible to full positive definite matrices (Uhler, 2012). Hence, the maximum likelihood estimator for $\boldsymbol{\Sigma}$ exists if and only if $\mathbf{Q}^{\mathcal{G}} \in \text{int}(\mathcal{C}(\mathbf{t}))$, which is equivalent to $\mathbf{Q}^{\mathcal{G}}$ being extensible to a full positive definite matrix. Whenever this happens, it is the only extensible matrix $\hat{\boldsymbol{\Sigma}}$ also satisfying the model restriction $\hat{\boldsymbol{\Sigma}}^{-1} \in \mathbb{S}^{>0}(\mathcal{G})$. Note that a sufficient condition is that $\mathbf{Q} \in \mathbb{S}^{>0}$, which almost surely holds for $N \geq p$.

The restriction in Equation (8), however, is not linear in the canonical parameter. In fact, Geiger and Meek (1998), Geiger et al. (2001) show that directed Gaussian Markov models form a curved exponential family. Thus, in the directed case, regular exponential family theory cannot

be applied to obtain the maximum likelihood estimators in a simplified way, which is why multivariate linear regression estimators is what is usually employed. Specifically, given an acyclic digraph $\mathcal{D} = (V, E)$, the \mathcal{D} -parameters for $(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ are estimated, for each $u \in V$, as the ordinary least squares estimators:

$$\begin{aligned}\hat{\boldsymbol{\beta}}_{u|\text{pa}(u)}^t &= \mathbf{q}_{u|\text{pa}(u)} \mathbf{Q}_{\text{pa}(u)|\text{pa}(u)}^{-1}, \\ \hat{\xi}_u &= \bar{x}_u - \hat{\boldsymbol{\beta}}_{u|\text{pa}(u)}^t \bar{\mathbf{x}}_{\text{pa}(u)}, \\ N\hat{v}_{uu} &= q_{uu} - \hat{\boldsymbol{\beta}}_{u|\text{pa}(u)}^t \mathbf{q}_{u|\text{pa}(u)}.\end{aligned}$$

The maximum likelihood estimator for $(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ can then be directly obtained from its respective \mathcal{D} -parameter estimators (see Andersson and Perlman, 1998, for an algorithm). As opposed to the undirected case, $(\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\Sigma}})$ exist with probability one if and only if $N \geq p + \max \{|\text{pa}(u)| : u \in V\}$.

5 Model selection via hypothesis testing

The maximum likelihood estimators presented in Section 4 can be used to address the problem of model estimation in Gaussian Markov models. A statistical procedure is required to previously select the graph that will define the Markov model. For this purpose, two main approaches can be differentiated when using hypothesis testing: stepwise selection and multiple testing methods.

5.1 Multiple testing

Gaussian Markov models are defined in terms of partial correlation. When the model is directed and vertices are assumed to be ancestrally ordered, we have (Section 5)

$$X_u \perp\!\!\!\perp X_v \mid \mathbf{X}_{\text{pa}(u)} \iff \beta_{uv|\text{pa}(u) \cup \{v\}} = 0 \iff \rho_{uv|\text{pa}(u)} = 0,$$

for $u \in V$, $v \in \text{pr}(u) \setminus \text{pa}(u)$, where the last equivalence follows from standard theory of partial correlation in multivariate Gaussian distributions (Anderson, 2003). On the other hand, in an undirected Gaussian Markov model,

$$X_u \perp\!\!\!\perp X_v \mid \mathbf{X}_{V \setminus \{u, v\}} \iff \sigma_{uv|V \setminus \{u, v\}} \iff \rho_{uv|V \setminus \{u, v\}} = 0.$$

This means that we can select the graph, for the Gaussian Markov model, based on the results of testing the multiple hypotheses $H_0^{uv} : \rho_{uv|U} = 0$ (for $u, v \in V$ and $U = V \setminus \{u, v\}$ when the graph is undirected, $u \in V$, $v \in \text{pr}(u)$ and $U = \text{pr}(u) \setminus \{v\}$ when the graph is acyclic directed).

The maximum likelihood estimator of $\rho_{uv|U}$ is the sample partial correlation, $\hat{\rho}_{uv|U}$, whose distribution is the same as that of the sample correlation but with restricted degrees of freedom. Specifically, if $\hat{\rho}_{uv}$ is the sample correlation between X_u and X_v , then, assuming $\rho_{uv} = 0$ (independence), $\sqrt{N-2} \hat{\rho}_{uv} / \sqrt{1 - \hat{\rho}_{uv}^2}$ is distributed as a Student's t with $N-2$ degrees of freedom. Now, if we consider the sample partial correlation and assume $\rho_{uv|U} = 0$ (conditional independence), then $\sqrt{N-|U|-2} \hat{\rho}_{uv|U} / \sqrt{1 - \hat{\rho}_{uv|U}^2}$ has a Student's t distribution with $N - |U| - 2$ degrees of freedom. As N increases, the distribution of $\hat{\rho}_{uv|U}$ tends to a Gaussian; however, a more accurate approximation can be obtained using Fisher's Z -transform,

$$Z(x) = \frac{1}{2} \log \left(\frac{1+x}{1-x} \right) = \tanh^{-1}(x).$$

The distribution of $Z(\hat{\rho}_{uv|U})$ tends to normality faster, and has a variance approximately independent of $\rho_{uv|U}$ (Anderson, 2003),

$$\sqrt{N - |U| - 3}(Z(\hat{\rho}_{uv|U}) - Z(\rho_{uv|U})) \xrightarrow{d} \mathcal{N}(0, 1).$$

The p -values obtained from testing each H_0^{uv} can be adjusted for controlling different error rates, see Drton and Perlman (2007) for a review on this topic.

5.2 Stepwise selection

An alternative to the previous multiple testing procedure is to test the hypothesis $H_0 : \boldsymbol{\Omega} \in \mathbb{S}^{\succ 0}(\mathcal{G}_0)$ against $H_1 : \boldsymbol{\Omega} \in \mathbb{S}^{\succ 0}(\mathcal{G})$, where $\mathcal{G}_0 = (V, E_{\mathcal{G}_0})$ is a subgraph of $\mathcal{G} = (V, E_{\mathcal{G}})$. The result of this test determines whether the edges in $E_{\mathcal{G}} \setminus E_{\mathcal{G}_0}$ should be excluded from the selected model. This is why these tests are usually called *edge exclusion tests* or *backward* model selection. Let $\hat{\boldsymbol{\Sigma}}_0$ and $\hat{\boldsymbol{\Sigma}}$ be the maximum likelihood estimators for a covariance matrix in the Markov model determined by \mathcal{G}_0 and \mathcal{G} , respectively. The likelihood ratio statistic is

$$T_L = \frac{\det(\hat{\boldsymbol{\Sigma}})^{\frac{N}{2}}}{\det(\hat{\boldsymbol{\Sigma}}_0)^{\frac{N}{2}}}.$$

Under H_0 , $-2 \log(T_L)$ is asymptotically distributed as $\chi^2_{|E_{\mathcal{G}}| - |E_{\mathcal{G}_0}|}$. In the undirected case, when \mathcal{G}_0 and \mathcal{G} have the same non-chordal minimal subgraphs, a better approximation than the χ^2 has been derived by Porteous (1989); Eriksen (1996). Specifically, if \mathcal{G} and \mathcal{G}_0 are chordal, then, under H_0 , $T_L^{2/N}$ is distributed as the product of univariate Beta variables parametrized according to a sequence of edge deletions from \mathcal{G} to \mathcal{G}_0 . In the directed case, the moments of the exact distribution of T_L have been obtained by Andersson and Perlman (1998).

6 Regularization

Regularization approaches (Bickel and Li, 2006), which simultaneously perform model selection and estimation, have become popular in the context of Markov models. They are usually applied when $N < p$, and thus the existence of the maximum likelihood estimator is not guaranteed.

Let X be a random vector whose distribution $\mathcal{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ belongs to an undirected Gaussian Markov model $\mathcal{N}(\mathcal{G})$, with $\mathcal{G} = (V, E)$. Since for each $u, v \in V$, $\beta_{uv|V \setminus \{u\}} = -\omega_{uv}/\omega_{uu}$ (Equation (3)), then

$$X_u \perp\!\!\!\perp X_v \mid \mathbf{X}_{V \setminus \{u, v\}} \iff \omega_{uv} = 0 \iff \beta_{uv|V \setminus \{u\}} = \beta_{vu|V \setminus \{v\}} = 0.$$

This means that an analogue of the matrix \mathbf{B} in directed Gaussian Markov models (Equation (7)) can be used for determining the missing edges in the undirected case. Specifically, let \mathbf{b}_u denote for $u \in V$ the vector such that $b_{uu} = 0$ and $b_{uv} = \beta_{uv|V \setminus \{u\}}$ when $v \neq u$. Assume, for notational simplicity, that $\boldsymbol{\mu} = \mathbf{0}$, and let $\mathbf{X} = \mathbf{x}$ be a $p \times N$ random sample from $\mathcal{N}_p(\mathbf{0}, \boldsymbol{\Sigma})$. Then, \mathbf{b}_u can be estimated as the solution to an optimization problem,

$$\hat{\mathbf{b}}_u^\lambda := \arg \min_{\mathbf{b}_u \in \mathbb{R}^p, b_{uu}=0} \left(\frac{1}{N} \|\mathbf{x}_{u\bullet} - \mathbf{x}^t \mathbf{b}_u\|_2^2 + \lambda f(\mathbf{b}_u) \right), \quad (9)$$

parametrized by a given $\lambda \geq 0$ and where $\mathbf{x}_{u\bullet}$ is the u -th row vector of \mathbf{x} and $f(\cdot)$ is the penalty function.

Observe that while $b_{uv} = 0 \iff b_{vu} = 0$, this does not necessarily hold for \hat{b}_{uv}^λ and \hat{b}_{vu}^λ . Hence, two different estimators for the edge set E may be defined: $\hat{E}_\wedge = \{uv : \hat{b}_{uv}^\lambda \neq 0 \text{ and } \hat{b}_{vu}^\lambda \neq 0\}$ and $\hat{E}_\vee = \{uv : \hat{b}_{uv}^\lambda \neq 0 \text{ or } \hat{b}_{vu}^\lambda \neq 0\}$. When $f(\cdot) = \|\cdot\|_1$, commonly known as the *lasso* penalty (Tibshirani, 1996) or l_1 regularization, both \hat{E}_\wedge and \hat{E}_\vee are consistent estimators of E for certain choice of λ . This result was independently discovered by Meinshausen and Bühlmann (2006), Zhao and Yu (2006), Zou (2006) and Yuan and Lin (2007b), and its sample complexity was thoroughly analysed by Wainwright (2009). However, the almost necessary and sufficient condition for such consistency, commonly called the ‘irrepresentable condition’, is rather restrictive. Hence, some variants have been proposed, which rely on thresholding \mathbf{b}_u or adding weights in the l_1 penalty, that under milder assumptions still achieve model selection consistency (Meinshausen and Yu, 2009) or other attractive, ‘oracle’ properties (van de Geer and Bühlmann, 2009). Bühlmann and van de Geer (2011), §7, review these alternatives.

In Gaussian directed Markov models, the correspondence between conditional independences and regression coefficients is reflected in the decomposition of covariance matrices: if $\mathcal{D} = (V, A)$ is an acyclic digraph and $\mathcal{N}_p(\mathbf{0}, \mathbf{\Sigma}) \in \mathcal{N}(\mathcal{D})$, then $\mathbf{\Omega} = (\mathbf{I}_p - \mathbf{B})^t \mathbf{V}^{-1} (\mathbf{I}_p - \mathbf{B})$ for some $\mathbf{V} \in \mathbf{\Delta}_p$ and $\mathbf{B} \in \mathbf{M}(\mathcal{D})$ (Equation (8)). Thus, the optimization problem in this case becomes

$$\hat{\mathbf{\Omega}}^\lambda = \arg \min_{\mathbf{\Omega} \in \mathbb{S}^{>0}(\mathcal{D}), \mathcal{D}=(V,A)} (\text{tr}(\mathbf{\Omega}\mathbf{Q}) - N \log \det(\mathbf{\Omega}) + \lambda f(\mathbf{\Omega})), \quad (10)$$

When $f(\mathbf{\Omega}) = |\{(u, v) : b_{uv} \neq 0\}| = |A|$, also called l_0 regularization, $\hat{\mathbf{\Omega}}^\lambda$ is a consistent estimator of $\mathbf{\Omega}$ for certain choice of λ (van de Geer and Bühlmann, 2013). Since the penalization is directly performed on the regression coefficients, this method is in certain sense related to the optimization problem in Equation (9) for undirected graphs. In fact, the assumptions required for the consistency of both methods share some symmetry, as we have outlined in Table 2. Note that one of the assumptions in van de Geer and Bühlmann (2013) involves permutations of the variables, this is because an ancestral order is assumed to be unknown. By contrast, Shojaie and Michailidis (2010) prove model selection consistency when such ordering is known, using a variation of lasso in (10).

Meinshausen and Bühlmann (2006)	van de Geer and Bühlmann (2013)
Lower bound on $ \rho_{uv V \setminus \{u,v\}} $	Lower bound on $ \beta_{uv \text{pr}(u)} $
Upper bound on $ \text{ne}(v) $	Upper bound on $ \text{pa}(v) $
Bounded neighbourhood perturbations	Bounded permutation perturbations

Table 2: Comparison of some of the assumptions for consistent model selection in Gaussian Markov models.

The method of Equation (10), usually called penalized likelihood method, can also be applied in the undirected case. Let $\|\mathbf{M}\|_{q+r}$ denote the $q + r$ -norm of the vectorized form of a $q \times r$ matrix \mathbf{M} , and \mathbf{M}^Δ the matrix with off-diagonal elements equal to the respective entries in \mathbf{M} and zero diagonal. Yuan and Lin (2007a) were the first to pursue this approach, and they chose $f(\mathbf{\Omega}) = \|\mathbf{\Omega}^\Delta\|_1$, that is, the off-diagonal elements in $\mathbf{\Omega}$, which determine the edges of the resulting undirected graph, are penalized. An alternative l_1 regularization was proposed in Banerjee et al. (2008), where the diagonal elements of $\mathbf{\Omega}$ were included, that is $f(\mathbf{\Omega}) = \|\mathbf{\Omega}\|_1$. However, since $1/\omega_{uu} = \sigma_{uu|V \setminus \{u\}}$, this choice for the penalty favours larger values for the error variances in the regression of X_u on the rest of variables (Bühlmann and van de Geer, 2011). Nonetheless, this estimator is the one chosen in the extensively used algorithm *Graphical Lasso* of Friedman et al. (2008), although model selection consistency has only been proved for

$f(\Omega) = \|\Omega^\Delta\|_1$ (Lam and Fan, 2009; Ravikumar et al., 2011). The sufficient conditions required for this consistency share some similarity with those for problem (10), although it is not known whether they are strictly stronger than the irrerepresentable condition for problem (9), as some examples (Meinshausen, 2008) seem to indicate.

The estimators obtained from Equations (9) and (10) allow to recover an estimator $\hat{\Sigma}^\lambda$ of Σ . Under the assumptions discussed, they further provide an estimator $\hat{\mathcal{G}}^\lambda$ of the undirected graph that consistently selects the underlying Gaussian Markov model. Hence, $\hat{\Sigma}^\lambda$ could be used for standard maximum likelihood estimation (Section 4) in $\mathcal{N}(\hat{\mathcal{G}}^\lambda)$, obtaining an alternative estimator of Σ . Zhou et al. (2011) provide convergence rates of one such two-step procedure, where method (9) is used as the first step.

7 Bayesian model selection and estimation

Consider a continuous multivariate family \mathcal{F}_θ parametrized by θ , and denote as $f(\mathbf{x} \mid \theta)$ the density function of a random sample \mathbf{X} from $P \in \mathcal{F}_\theta$ for a given value of θ . In Bayesian statistics, θ is treated as a random variable with known distribution, $\mathcal{L}(\theta)$, usually called the *prior* distribution of θ . Inference is then performed based on the value of $\mathcal{L}(\theta \mid \mathbf{x}) \propto \mathcal{L}(\theta)P(\mathbf{x} \mid \theta)$, the *posterior* distribution of θ given the information in $\mathbf{X} = \mathbf{x}$.

Dawid and Lauritzen (1993) characterized prior distributions in terms of properties of the graph associated with the Markov model, mimicking Markov properties. Specifically, let θ denote the parameters of a distribution $P_{\mathbf{X}} \in \mathcal{M}(\mathcal{G})$ with \mathcal{G} chordal, and for subsets $A, B \subseteq V$, denote as θ_A and $\theta_{B|A}$ the parameters of the marginal distribution of \mathbf{X}_A and the conditional distribution of \mathbf{X}_A given values of \mathbf{X}_B , respectively. $\mathcal{L}(\theta)$ is said to be

- *weak hyper \mathcal{G} - Markov* if $\theta_{A \cup S} \perp\!\!\!\perp \theta_{B \cup S} \mid \theta_S$ for any decomposition (A, B, S) of \mathcal{G} ,
- *strong hyper \mathcal{G} -Markov* if $\theta_{B \cup S|A \cup S} \perp\!\!\!\perp \theta_{A \cup S}$ for any decomposition (A, B, S) of \mathcal{G} .

Strong hyper Markov prior distributions allow to localize computations over the graph cliques when performing Bayesian inference: if $\mathcal{L}(\theta)$ is strong hyper \mathcal{G} -Markov and $\mathfrak{C}(\mathcal{G})$ is the collection of cliques in \mathcal{G} , then $\mathcal{L}(\theta \mid \mathbf{x})$ is also strong hyper \mathcal{G} -Markov and $\mathcal{L}(\theta_C \mid \mathbf{x}) = \mathcal{L}(\theta_C \mid \mathbf{x}_C)$ for all $C \in \mathfrak{C}(\mathcal{G})$, where \mathbf{x}_C stands for all the observations in \mathbf{x} corresponding to the variable in C , that is, the matrix of row vectors from \mathbf{x} indexed by C .

Let $\mathcal{N}(\mathcal{G})$ be an undirected Gaussian Markov model with \mathcal{G} chordal, and $\mathcal{N}_p(\mathbf{0}, \Sigma) \in \mathcal{N}(\mathcal{G})$. When \mathcal{G} is complete, that is, under no constraint on Ω , the inverse Wishart, denoted as $\mathcal{W}_p^{-1}(\nu, \Psi)$ with $\nu \in \mathbb{R}$, $\nu > p - 1$ and $\Psi \in \mathbb{M}_{p \times p}(\mathbb{R})$, $\Psi \succ 0$, is a conjugate prior for Σ (Anderson, 2003). This result translates to hyper Markov distributions for a general \mathcal{G} : setting $\mathcal{L}(\Sigma_{CC}) = \mathcal{W}_{|C|}^{-1}(\nu, \Psi^C)$ for each $C \in \mathfrak{C}(\mathcal{G})$, there exists a unique strong hyper \mathcal{G} -Markov distribution with such marginals (Dawid and Lauritzen, 1993). Such prior distribution is called the *hyper inverse Wishart distribution* and denoted by $\mathcal{HW}_p^{-1}(\nu, \Psi)$, where $\Psi \in \mathbb{S}^{>0}$ and $\Psi_{CC} = \Psi^C$ for each $C \in \mathfrak{C}(\mathcal{G})$.

Since its introduction, the hyper inverse Wishart prior distribution has been extensively studied. The explicit expression for its density is given by Roverato (2000). Furthermore, since the absent edges of \mathcal{G} correspond to zeros in Ω (Equation (4)), Roverato (2000) derives the distribution induced on Ω by assuming $\mathcal{L}(\Sigma) = \mathcal{HW}_p^{-1}(\nu, \Psi)$. He shows that in that case, the density is proportional to that of a Wishart matrix conditioned on the event $\Omega \in \mathbb{S}^{>0}(\mathcal{G})$, and calls such prior distribution on Ω the *\mathcal{G} -conditional Wishart*. Letac and Massam (2007) generalize the \mathcal{G} -conditional Wishart to a broader conjugate family of Wishart distributions, which they call *Type II Wisharts*, sharing similar properties. Hyper Markov priors have also been defined when

\mathcal{G} is an acyclic digraph (Dawid and Lauritzen, 1993), and extended in Letac and Massam (2007) to *Type I Wishart* distributions. Closed form expressions of the Bayes estimators associated with both Type I and II Wishart priors have been derived by Rajaratnam et al. (2008).

While hyper Markov distributions have only been defined for chordal graphs, the hyper inverse Wishart has been extended to the non-chordal case by Roverato (2002), based on properties of the Isserlis matrix of Σ (Roverato and Whittaker, 1998). Hyper Markov distributions have also been applied for setting $\mathcal{L}(\mathcal{G})$, the prior distribution for model selection, by Byrne and Dawid (2015), for both undirected and acyclic directed graphs. Other common choices for $\mathcal{L}(\mathcal{G})$, when \mathcal{G} is undirected are the uniform distribution, which is biased towards medium-sized graphs, or prior distributions that favour sparse graphs (Jones et al., 2005).

The methodology by Geiger and Heckerman (2002) can be seen as an extension of hyper Markov priors for acyclic digraphs, since they both coincide when for chordal skeletons. Specifically, if $\mathcal{D} = (V, E)$ is an acyclic digraph, the posterior density of \mathcal{D} is $\pi(\mathcal{D} \mid \mathbf{x}) \propto \pi(\mathcal{D}) f(\mathbf{x} \mid \mathcal{D}) = \pi(\mathcal{D}) \int f(\mathbf{x} \mid \mathcal{D}, \boldsymbol{\theta}) \pi(\boldsymbol{\theta} \mid \mathcal{D}) d\boldsymbol{\theta}$. Under some assumptions on $\pi(\boldsymbol{\theta} \mid \mathcal{D})$ and $f(\mathbf{x} \mid \mathcal{D}, \boldsymbol{\theta})$, computations can be localized as

$$f(\mathbf{x} \mid \mathcal{D}) = \prod_{v \in V} \frac{f(\mathbf{x}_{\{v\} \cup \text{pa}_{\mathcal{D}}(v)} \mid \mathcal{D}_c)}{f(\mathbf{x}_{\text{pa}_{\mathcal{D}}(v)} \mid \mathcal{D}_c)}, \quad (11)$$

where \mathcal{D}_c is an arbitrary complete digraph with vertex set V .

When the parametric family is the multivariate Gaussian, the conjugate prior for $\boldsymbol{\theta} = (\boldsymbol{\mu}, \boldsymbol{\Omega})$ is the Gaussian-Wishart distribution. In fact, Geiger and Heckerman (2002) identify the Gaussian-Wishart as the only prior satisfying the *global parameter independence* assumption, that is, $\mathcal{L}(\boldsymbol{\theta} \mid \mathcal{D}) = \prod_{v \in V} \mathcal{L}(\boldsymbol{\theta}_v \mid \mathcal{D})$, which is required for Equation (11) to hold. This scope of prior distributions can be widened when an ancestral ordering of V is assumed fixed, as analysed in Roverato (2002).

8 Open problems and current research

Directed and undirected Markov models are traditional models, and thus a plethora of methodological developments are available, as we have shown. Despite this, several open problems remain that deserve further research.

At the foundational level, independence relations can be generalised to what are known as *separoids* (Dawid, 2001), a relaxation of graphoids. Separoids usually appear whenever a notion of ‘irrelevance’ is being mathematically treated, although they are sometimes too restrictive (Cozman and Walley, 2005). Further research on these axiom systems from an abstract point of view could shed more light on how the apparently different mathematical contexts in which such structures arise are related, and also provide an explicit bridge between them and the recently defined *independence logic* (Grädel and Väänänen, 2013), closely related.

Hammersley and Clifford’s theorem (Hammersley and Clifford, 1971) is a straightforward tool for checking whether an independence model originated from a distribution is a graphoid; however, the condition is not necessary and sufficient. Further relaxations remain to be found, although positivity is generally essential (Moussouris, 1974). Also regarding independence models induced by graphs, recall the space of Markov equivalence classes of acyclic digraphs, \mathfrak{D}_p / \sim . The asymptotic ratio $l = \lim_{p \rightarrow \infty} |\mathfrak{D}_p| / |\mathfrak{D}_p / \sim|$ influences the computational gain obtained by using \mathfrak{D}_p / \sim instead of \mathfrak{D}_p as a search space for model selection. Steinsky (2004) analytically calculates an upper bound of l as 13.65. Exact computations by Gillispie and Perlman (2002), for $p \leq 10$, and approximations by Sonntag et al. (2015), up to $p = 31$, seem to indicate that $l \sim 3.7$. However, its analytical deduction remains an open problem. Note that the computational gain

is not only influenced by l , but also by other factors, such as how the element size in \mathfrak{D}_p/\sim is distributed.

The existence of the maximum likelihood estimator in Gaussian undirected Markov models has not yet been completely characterized, although some progress has been made. Since the problem lies at the interface between statistics and linear algebra, several of the existing results have been independently discovered by researchers in both areas. For chordal graphs, the problem was solved separately by Grone et al. (1984) and Frydenberg and Lauritzen (1989): if $\mathfrak{C}(\mathcal{G}^*)$ is the class of cliques in \mathcal{G} , \mathcal{G}^* is a chordal cover of \mathcal{G} , and $q^* := \max\{|C| : C \in \mathfrak{C}(\mathcal{G}^*)\}$, $q := \max\{|C| : C \in \mathfrak{C}(\mathcal{G})\}$; then there is a probability of one that $\widehat{\Sigma}$ exists if $N \geq q^*$, and does not exist if $N < q$. This result does not account for the case $q \leq N < q^*$ for non-chordal graphs, since otherwise $q = q^*$. Special subtypes of non-chordal graphs are p -cycles, which were addressed by Barrett et al. (1993) (from the viewpoint of linear algebra), and separately by Buhl (1993) (from the perspective of statistics): there is a probability strictly between zero and one of $\widehat{\Sigma}$ existing if $N = 2$. Uhler (2012) recently detailed, from the algebraic viewpoint, results paralleling findings reported by Buhl (1993) for bipartite graphs.

Regarding hypothesis testing, a backward stepwise method has become popular for estimating the acyclic digraph $\mathcal{D} = (V, E)$ of a Markov model $\mathbf{M}(\mathcal{D})$, commonly called the *PC algorithm* (Spirtes et al., 2000). This method proceeds by first estimating the skeleton, \mathcal{D}^U , of \mathcal{D} from a complete undirected graph, and then orienting it. Note that this algorithm can also be used for undirected Markov models, omitting the second orientation step. At iteration $i \in \{1, \dots, |V|\}$ of the first step, $H_0 : X_u \perp\!\!\!\perp X_v \mid X_C$ is tested, with $|C| = i - 1$ and $P_{\mathbf{X}} \in \mathbf{M}(\mathcal{D})$, and the edge uv is removed from $\widehat{\mathcal{D}}^U$ if H_0 is not rejected. $\widehat{\mathcal{D}}^U$ depends on the order in which H_0 is tested at each iteration, problem circumvented in the modification by Colombo and Maathuis (2014). Assuming that $\mathcal{I}(\mathbf{X}) = \mathcal{I}(\mathcal{D})$ (see Section 2), commonly called the *faithfulness* assumption, Robins et al. (2003) showed that the PC algorithm is pointwise consistent but may not be uniformly consistent, regardless of the method used for testing H_0 . Zhang and Spirtes (2003) approached this problem by introducing a stronger condition, called *strong faithfulness*, which gives uniform consistency, even in a high-dimensional setting (Kalisch and Bühlmann, 2007). However, despite the set of ‘unfaithful’ distributions has Lebesgue measure zero (Meek, 1995), those ‘strongly unfaithful’ constitute a non-zero Lebesgue measure set, which can in some cases be very large (Uhler et al., 2013). In the Gaussian case this is related with the assumption of bounded partial correlations. The characteristics of the geometric surface determined by such assumption is currently being researched (Lin et al., 2014).

The bounding of partial correlations assumed in the PC algorithm resembles the assumptions used in regularization methods (Table 2). In fact, l_0 regularization (van de Geer and Bühlmann, 2013) has been suggested as an alternative for the PC, in order to avoid the restrictive strong faithfulness assumption. However, it is unclear how the assumptions of both methods are related (Uhler et al., 2013). Noticeably, regularization techniques have been more developed for undirected Gaussian Markov models, since the problem becomes significantly easier thanks to $\Omega \in \mathbb{S}^{\succ 0}(\mathcal{G})$ in Equation (10) being a linear constraint when \mathcal{G} is undirected. Recently, Aragam and Zhou (2015) have proposed an algorithm to calculate regularized estimators from Equation (10) when the penalty function is concave.

In the Bayesian approach to Markov models, one of the main difficulties still encountered is to compute the normalizing constant of the hyper inverse Wishart distribution in the non-chordal case, which does not have a closed-form expression (Roverato, 2002). Atay-Kayis and Massam (2005) analysed the Cholesky decomposition of Ω and its relation with the cone $\mathbb{S}^{\succ 0}(\mathcal{G})$ and extensible projections (Section 4). Carvalho et al. (2007) and Wang and Carvalho (2010) used such theoretical analysis to provide a direct sampler from the hyper inverse Wishart prior. Recently, exact formulas for these constants seem to have been found by Uhler et al. (2016).

A Graph Theory

A graph is defined as a pair $\mathcal{G} = (V, E)$ where V is the set of vertices and E is the set of edges. Throughout the whole paper, the graphs will, unless otherwise stated, be labelled and simple. This means that the elements in V are labelled, for example, as $1, \dots, p$; and E is composed of pairs of distinct elements in V . A graph is called *undirected* if these pairs are unordered ($E \subseteq \{\{u, v\} : u, v \in V\}$), and *directed* or *digraph* otherwise ($E \subseteq \{(u, v) : u, v \in V\}$). Edges $\{u, v\}$ in an undirected graph are usually denoted by uv and drawn as a line (see Figure 3a). In a digraph, however, they are called *arcs* or *directed edges* and represented as arrows (Figure 3b and 3c).

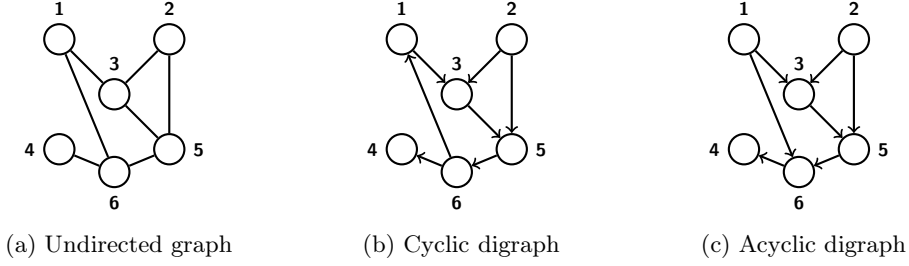


Figure 3: Examples of an undirected graph and two digraphs.

A.1 Undirected graphs

In an undirected graph $\mathcal{G} = (V, E)$, if $uv \in E$, u and v are called *neighbours*. For $v \in V$, the set of its neighbours is denoted by $\text{ne}(v)$, and the *closure* of v is $\text{cl}(v) := \{v\} \cup \text{ne}(v)$. \mathcal{G} is called *complete* if for every $u, v \in V$, $uv \in E$. A maximal $C \subseteq V$ such that \mathcal{G}_C is complete is called a *clique*. Let $\mathcal{H} = (V_{\mathcal{H}}, E_{\mathcal{H}})$ be another undirected graph. \mathcal{H} is a *subgraph* of \mathcal{G} (written as $\mathcal{H} \subseteq \mathcal{G}$) if $V_{\mathcal{H}} \subseteq V$ and $E_{\mathcal{H}} \subseteq E$. If $E_{\mathcal{H}} = \{uv \in E : u, v \in V_{\mathcal{H}}\}$, then \mathcal{H} is called the *induced subgraph* and denoted by $\mathcal{G}_{V_{\mathcal{H}}}$.

A *walk* between u and v is an ordered sequence of vertices u_0, \dots, u_k , where $u_0 = u$, $u_k = v$ and $u_{i-1}u_i \in E$ for $i \in \{1, \dots, k\}$. The number k is called the *length* of the walk. If $u = v$, the walk is *closed*, and when u_0, \dots, u_{k-1} are distinct, the walk is called a *path*. A closed path of length $k \geq 3$ is called a *cycle* or k -cycle. \mathcal{G} is called *chordal* or *triangulated* if all minimal k -cycles are of length $k = 3$. A *chordal cover* of a graph \mathcal{G} is a graph \mathcal{G}^* such that $\mathcal{G} \subseteq \mathcal{G}^*$ and \mathcal{G}^* is chordal.

$S \subseteq V$ *separates* u and v in $\mathcal{G} = (V, E)$ if there is no path between u and v in the subgraph $\mathcal{G}_{V \setminus S}$. If we consider $A, B, S \subseteq V$, A and B are said to be *separated* by S if u and v are separated by S for all $u \in A$, $v \in B$. Let V be partitioned into disjoint sets $A, B, S \subseteq V$. (A, B, S) is called a *decomposition* of \mathcal{G} if S separates A and B in \mathcal{G} and \mathcal{G}_S is complete. If $A \neq \emptyset$ and $B \neq \emptyset$, the decomposition is said to be *proper*. An undirected graph is *decomposable* if: (i) it is complete or (ii) it admits a proper decomposition into *decomposable* subgraphs. An undirected graph is decomposable if and only if it is chordal.

A.2 Acyclic digraphs

In a digraph $\mathcal{D} = (V, A)$, the definitions of (induced) subgraph, walk, path, and cycle are analogous to the undirected case. The undirected graph $\mathcal{D}^U := (V, A^U)$ with $A^U := \{uv : (u, v) \in A\}$

is called the *skeleton* of \mathcal{D} , and \mathcal{D} is one of its *orientations*. A digraph \mathcal{D} is said to be *complete* when \mathcal{D}^U is complete.

In the following, assume that \mathcal{D} is acyclic (see Figure 3b and Figure 3c for a cyclic and an acyclic digraph, respectively). The *parent set* of $v \in V$ is $\text{pa}(v) := \{u \in V : (u, v) \in A\}$. Conversely, the *child set* is $\text{ch}(v) := \{u \in V : (v, u) \in A\}$. The *ancestors* of v , $\text{an}(v)$, are those $u \in V$ such that there exists a directed path from u to v ; the *descendants* of v , $\text{de}(v)$, are those $u \in V$ such that there exists a directed path from v to u . The set $\text{nd}(v) := V \setminus (\{v\} \cup \text{de}(v))$ will be the set of *non-descendants* of $v \in V$, and $\text{An}(A) := A \cup (\cup_{a \in A} \text{an}(a))$ the *ancestral set* of $A \subseteq V$. Note that a total order \prec can be defined over the set of vertices V in an acyclic digraph $\mathcal{D} = (V, A)$, such that if $(u, v) \in A$, then $u \prec v$. This ordering is usually called *ancestral*, and it is a linear extension of the partial order naturally defined as $u \preceq v$ if $u \in \text{an}(v)$. For $v \in V$, the set of *successors* of v with respect to \prec is $\text{su}(v) = \{u \in V : u \succ v\}$; the set of *predecessors* of v is $\text{pr}(v) = \{u \in V : u \prec v\}$.

Finally, let $u, w_1, w_2 \in V$ with $(w_1, u), (w_2, u) \in A$ and $(w_1, w_2), (w_2, w_1) \notin A$ (see vertices 1, 2 and 3 in Figure 3c). Such configurations are usually called *v-structures* and denoted by $w_1 \rightarrow u \leftarrow w_2$. The *moral graph* of \mathcal{D} is defined as the undirected graph \mathcal{D}^m with edge set $A^U \cup \{w_1 w_2 : w_1 \rightarrow u \leftarrow w_2 \text{ for some } u \in V\}$.

References

- Anderson, T. W. (2003). *An Introduction to Multivariate Statistical Analysis*. John Wiley & Sons, 3rd edition.
- Andersson, S. A. and Perlman, M. D. (1998). Normal linear regression models with recursive graphical Markov structure. *Journal of Multivariate Analysis*, 66(2):133 – 187.
- Aragam, B. and Zhou, Q. (2015). Concave penalized estimation of sparse gaussian bayesian networks. *Journal of Machine Learning Research*, 16:2273–2328.
- Atay-Kayis, A. and Massam, H. (2005). A monte carlo method for computing the marginal likelihood in nondecomposable gaussian graphical models. *Biometrika*, 92(2):317–335.
- Banerjee, O., El Ghaoui, L., and d’Aspremont, A. (2008). Model selection through sparse maximum likelihood estimation for multivariate Gaussian or binary data. *Journal of Machine Learning Research*, 9:485–516.
- Barndorff-Nielsen, O. (1978). *Information and Exponential Families in Statistical Theory*. John Wiley & Sons.
- Barrett, W., Johnson, C. R., and Tarazaga, P. (1993). The real positive definite completion problem for a simple cycle. *Linear Algebra and Its Applications*, 192:3 – 31.
- Besag, J. (1974). Spatial interaction and the statistical analysis of lattice systems. *J. R. Statist. Soc. B*, 36(2):192–236.
- Bühlmann, P. and van de Geer, S. (2011). *Statistics for High-Dimensional Data: Methods, Theory and Applications*. Springer.
- Bickel, P. J. and Li, B. (2006). Regularization in statistics. *Test*, 15(2):271–344.
- Buhl, S. L. (1993). On the existence of maximum likelihood estimators for graphical Gaussian models. *Scand. J. Statist.*, 20(3):263–270.

- Byrne, S. and Dawid, A. P. (2015). Structural markov graph laws for bayesian model uncertainty. *Ann. Statist.*, 43(4):1647–1681.
- Carvalho, C. M., Massam, H., and West, M. (2007). Simulation of hyper-inverse wishart distributions in graphical models. *Biometrika*, 94(3):647–659.
- Colombo, D. and Maathuis, M. H. (2014). Order-independent constraint-based causal structure learning. *Journal of Machine Learning Research*, 15:3921–3962.
- Cox, D. R. and Wermuth, N. (1993). Linear dependencies represented by chain graphs. *Statistical Science*, 8(3):204–218.
- Cozman, F. G. and Walley, P. (2005). Graphoid properties of epistemic irrelevance and independence. *Annals of Mathematics and Artificial Intelligence*, 45(1):173–195.
- Daly, R., Shen, Q., and Aitken, S. (2011). Learning Bayesian networks: Approaches and issues. *The Knowledge Engineering Review*, 26:99–157.
- Darroch, J. N., Lauritzen, S. L., and Speed, T. P. (1980). Markov fields and log-linear interaction models for contingency tables. *Ann. Statist.*, 8(3):522–539.
- Dawid, A. P. (1979). Conditional independence in statistical theory. *J. R. Statist. Soc. B*, 41(1):1–31.
- Dawid, A. P. (1980). Conditional independence for statistical operations. *Ann. Statist.*, 8(3):598–617.
- Dawid, A. P. (2001). Separoids: A mathematical framework for conditional independence and irrelevance. *Annals of Mathematics and Artificial Intelligence*, 32(1):335–372.
- Dawid, A. P. and Lauritzen, S. L. (1993). Hyper Markov laws in the statistical analysis of decomposable graphical models. *Ann. Statist.*, 21(3):1272–1317.
- Dempster, A. P. (1972). Covariance selection. *Biometrics*, 28(1):157–175.
- Drton, M. and Perlman, M. D. (2007). Multiple testing and error control in Gaussian graphical model selection. *Statistical Science*, 22(3):430–449.
- Eriksen, P. S. (1996). Tests in covariance selection models. *Scand. J. Statist.*, 23(3):275–284.
- Friedman, J., Hastie, T., and Tibshirani, R. (2008). Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, 9(3):432–441.
- Frydenberg, M. (1990). The chain graph Markov property. *Scand. J. Statist.*, 17(4):333–353.
- Frydenberg, M. and Lauritzen, S. L. (1989). Decomposition of maximum likelihood in mixed graphical interaction models. *Biometrika*, 76(3):539–555.
- Geiger, D. and Heckerman, D. (2002). Parameter priors for directed acyclic graphical models and the characterization of several probability distributions. *Ann. Statist.*, 30(5):1412–1440.
- Geiger, D., Heckerman, D., King, H., and Meek, C. (2001). Stratified exponential families: Graphical models and model selection. *Ann. Statist.*, 29(2):505–529.
- Geiger, D. and Meek, C. (1998). Graphical models and exponential families. In *Proceedings of the Fourteenth Conference on Uncertainty in Artificial Intelligence*, UAI’98, pages 156–165, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

- Geiger, D. and Pearl, J. (1990). On the logic of causal models. In *Proc. of the Fourth Annual Conference on Uncertainty in Artificial Intelligence*, pages 3–14, Corvallis. AUAI Press.
- Geiger, D. and Pearl, J. (1993). Logical and algorithmic properties of conditional independence and graphical models. *Ann. Statist.*, 21(4):2001–2021.
- Gillispie, S. B. and Perlman, M. D. (2002). The size distribution for Markov equivalence classes of acyclic digraph models. *Artificial Intelligence*, 141(1–2):137 – 155.
- Grädel, E. and Väänänen, J. (2013). Dependence and independence. *Studia Logica*, 101(2):399–410.
- Grimmett, G. R. (1973). A theorem about random fields. *Bulletin of the London Mathematical Society*, 5(1):81–84.
- Grone, R., Johnson, C. R., Sá, E. M., and Wolkowicz, H. (1984). Positive definite completions of partial Hermitian matrices. *Linear Algebra and Its Applications*, 58:109 – 124.
- Hammersley, J. M. and Clifford, P. (1971). Markov fields on finite graphs and lattices. Unpublished manuscript.
- Horn, R. A. and Johnson, C. R. (2012). *Matrix Analysis*. Cambridge University Press, 2nd edition.
- Howard, R. A. and Matheson, J. E. (2005). Influence diagrams. *Decision Analysis*, 2(3):127–143.
- Ibáñez, A., Armañanzas, R., Bielza, C., and Larrañaga, P. (2016). Genetic algorithms and Gaussian Bayesian networks to uncover the predictive core set of bibliometric indices. *Journal of the Association for Information Science and Technology*, 67(7):1703–1721.
- Isham, V. (1981). An introduction to spatial point processes and Markov random fields. *International Statistical Review*, 49(1):21–43.
- Jones, B., Carvalho, C., Dobra, A., Hans, C., Carter, C., and West, M. (2005). Experiments in stochastic computation for high-dimensional graphical models. *Statistical Science*, 20(4):388–400.
- Kalisch, M. and Bühlmann, P. (2007). Estimating high-dimensional directed acyclic graphs with the PC-algorithm. *Journal of Machine Learning Research*, 8:613–636.
- Lam, C. and Fan, J. (2009). Sparsistency and rates of convergence in large covariance matrix estimators. *The Annals of Statistics*, 37(6B):4254–4278.
- Lauritzen, S. L. (1996). *Graphical Models*. Oxford University Press.
- Lauritzen, S. L., Dawid, A. P., Larsen, B. N., and Leimer, H.-G. (1990). Independence properties of directed Markov fields. *Networks*, 20(5):491–505.
- Letac, G. and Massam, H. (2007). Wishart distributions for decomposable graphs. *Ann. Statist.*, 35(3):1278–1323.
- Lin, S., Uhler, C., Sturmfels, B., and Bühlmann, P. (2014). Hypersurfaces and their singularities in partial correlation testing. *Foundations of Computational Mathematics*, 14(5):1079–1116.

- Meek, C. (1995). Strong completeness and faithfulness in bayesian networks. In *Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence*, UAI'95, pages 411–418, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.
- Meinshausen, N. (2008). A note on the lasso for gaussian graphical model selection. *Statistics & Probability Letters*, 78(7):880 – 884.
- Meinshausen, N. and Bühlmann, P. (2006). High-dimensional graphs and variable selection with the lasso. *Ann. Statist.*, 34(3):1436–1462.
- Meinshausen, N. and Yu, B. (2009). Lasso-type recovery of sparse representations for high-dimensional data. *Ann. Statist.*, 37(1):246–270.
- Moussouris, J. (1974). Gibbs and Markov random systems with constraints. *Journal of Statistical Physics*, 10(1):11–33.
- Pearl, J. (1985). Bayesian networks: A model of self-activated memory for evidential reasoning. Technical Report R-43, University of California, Los Angeles.
- Pearl, J. (1986). Fusion, propagation, and structuring in belief networks. *Artificial Intelligence*, 29(3):241 – 288.
- Pearl, J. (1988). *Probabilistic Reasoning in Intelligent Systems*. Morgan Kaufmann.
- Pearl, J. and Paz, A. (1987). Graphoids: A graph-based logic for reasoning about relevance relations. In *Advances in Artificial Intelligence*, volume 2, pages 357–363. Elsevier.
- Porteous, B. T. (1989). Stochastic inequalities relating a class of log-likelihood ratio statistics to their asymptotic χ^2 distribution. *Ann. Statist.*, 17(4):1723–1734.
- Rajaratnam, B., Massam, H., and Carvalho, C. M. (2008). Flexible covariance estimation in graphical Gaussian models. *Ann. Statist.*, 36(6):2818–2849.
- Ravikumar, P., Wainwright, M. J., Raskutti, G., and Yu, B. (2011). High-dimensional covariance estimation by minimizing l_1 -penalized log-determinant divergence. *Electronic Journal of Statistics*, 5:935–980.
- Robins, J. M., Scheines, R., Spirtes, P., and Wasserman, L. (2003). Uniform consistency in causal inference. *Biometrika*, 90(3):491–515.
- Roverato, A. (2000). Cholesky decomposition of a hyper inverse wishart matrix. *Biometrika*, 87(1):99–112.
- Roverato, A. (2002). Hyper inverse wishart distribution for non-decomposable graphs and its application to bayesian inference for gaussian graphical models. *Scandinavian Journal of Statistics*, 29(3):391–411.
- Roverato, A. and Whittaker, J. (1998). The Isserlis matrix and its application to non-decomposable graphical Gaussian models. *Biometrika*, 85(3):711–725.
- Shojaie, A. and Michailidis, G. (2010). Penalized likelihood methods for estimation of sparse high-dimensional directed acyclic graphs. *Biometrika*, 97(3):519–538.
- Sonntag, D., Peña, J. M., and Gómez-Olmedo, M. (2015). Approximate counting of graphical models via MCMC revisited. *International Journal of Intelligent Systems*, 30(3):384–420.

- Speed, T. P. (1979). A note on nearest-neighbour Gibbs and Markov probabilities. *Sankhya A*, 41(3/4):184–197.
- Spirites, P., Glymour, C., and Scheines, R. (2000). *Causation, Prediction, and Search*. MIT Press.
- Steinsky, B. (2004). Asymptotic behaviour of the number of labelled essential acyclic digraphs and labelled chain graphs. *Graphs and Combinatorics*, 20(3):399–411.
- Studeny, M. (2005). *On Probabilistic Conditional Independence Structures*. Springer London.
- Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *J. R. Statist. Soc. B*, 58(1):267–288.
- Uhler, C. (2012). Geometry of maximum likelihood estimation in Gaussian graphical models. *Ann. Statist.*, 40(1):238–261.
- Uhler, C., Lenkoski, A., and Richards, D. (2016). Exact formulas for the normalizing constants of wishart distributions for graphical models. *Submitted*. arXiv:1406.4901.
- Uhler, C., Raskutti, G., Bühlmann, P., and Yu, B. (2013). Geometry of the faithfulness assumption in causal inference. *Annals of Statistics*, 41(2):436–463.
- van de Geer, S. and Bühlmann, P. (2013). ℓ_0 -penalized maximum likelihood for sparse directed acyclic graphs. *Ann. Statist.*, 41(2):536–567.
- van de Geer, S. A. and Bühlmann, P. (2009). On the conditions used to prove oracle results for the lasso. *Electron. J. Statist.*, 3:1360–1392.
- Verma, T. and Pearl, J. (1991). Equivalence and synthesis of causal models. In *Proc. of the Sixth Annual Conference on Uncertainty in Artificial Intelligence*, pages 255–270, Corvallis. AUAI Press.
- Wainwright, M. J. (2009). Sharp thresholds for high-dimensional and noisy sparsity recovery using ℓ_1 -constrained quadratic programming (lasso). *IEEE Transactions on Information Theory*, 55(5):2183–2202.
- Wang, H. and Carvalho, C. M. (2010). Simulation of hyper-inverse wishart distributions for non-decomposable graphs. *Electron. J. Statist.*, 4:1470–1475.
- Werhli, A. V., Grzegorzcyk, M., and Husmeier, D. (2006). Comparative evaluation of reverse engineering gene regulatory networks with relevance networks, graphical Gaussian models and Bayesian networks. *Bioinformatics*, 22(20):2523–2531.
- Wermuth, N. (1976). Analogies between multiplicative models in contingency tables and covariance selection. *Biometrics*, 32(1):95–108.
- Wermuth, N. (1980). Linear recursive equations, covariance selection, and path analysis. *J. Am. Statist. Ass.*, 75(372):963–972.
- Wermuth, N. (2015). Graphical Markov models, unifying results and their interpretation. *Wiley StatsRef: Statistics Reference Online*, pages 1–29.
- Wermuth, N. and Lauritzen, S. L. (1983). Graphical and recursive models for contingency tables. *Biometrika*, 70(3):537–552.

- Wright, S. (1934). The method of path coefficients. *Ann. Math. Statist.*, 5(3):161–215.
- Yuan, M. and Lin, Y. (2007a). Model selection and estimation in the Gaussian graphical model. *Biometrika*, 94(1):19–35.
- Yuan, M. and Lin, Y. (2007b). On the non-negative garrotte estimator. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 69(2):143–161.
- Yule, G. U. (1907). On the theory of correlation for any number of variables, treated by a new system of notation. *Proc. R. Soc. A*, 79(529):182–193.
- Zhang, J. and Spirtes, P. (2003). Strong faithfulness and uniform consistency in causal inference. In *Proceedings of the Nineteenth Conference on Uncertainty in Artificial Intelligence (UAI-2003)*, pages 632–639. Morgan Kaufmann.
- Zhao, P. and Yu, B. (2006). On model selection consistency of lasso. *Journal of Machine Learning Research*, 7:2541–2563.
- Zhou, S., Rütimann, P., Xu, M., and Bühlmann, P. (2011). High-dimensional covariance estimation based on Gaussian graphical models. *Journal of Machine Learning Research*, 12:2975–3026.
- Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American Statistical Association*, 101(476):1418–1429.